Business Case

Assuming you are working in a bank and a customer is requesting for credit card.

You need to estimate the amount of credit you could grant to the customer or reject the application

Hence you need to predict if the customer will default or not.

You are to create model to predict the default using machine learning model (regression, decision tree, random forest, XGboost, SKlearn neural network)

Data visualization is needed. At least use one of pandas, seaborn and matplotlib.

All the above contribute to 50%.

The other 50% are:

How to improve the results through programming by changing the features selected, data wrangling or parameter setting (optional: you may explain how and why you are doing so or you could explain why the result CANNOT be improved).

Qualitatively, explain the pros and cons about all the models you use.

How to overcome the weakness of your all your models (future study).

Please answer:

1. Why is credit card default prediction important for financial institutions?

2. How do consumers benefit from credit card default prediction?

3. How does credit card default prediction contribute to economic stability?

4. What is the relationship between credit card default prediction and fraud detection?

5. How has technology like AI transformed credit card default prediction?

Hint:

You need about 70% to impute data.

For keras, input\_dim at layer one is len(X\_train.columns), 1st hidden layer neurons should be similar to the input layer, second hidden layer is optional. Output layer should have only 1 neurons, activation is "sigmoid" and should not have any dropout later.

For categorical imputation, please refer to notes (you could either impute with most frequently happen category or create a column for unknow category).

Try not to remove too many rows because of 1 columns is invalid as there are a lot of information from other columns.

You need to make assumption when needed. (please make reasonable assumption). Think extrapolation or interpolation.

If an invalid category happen too often, then most likely you need to make assumption (data dictionary may not be 100% complete).

You may put all your assumption and qualitative answers in python code.